DOCUMENT RESUME

ED 330 722 TM 016 318

AUTHOR Lunz, Mary E.; And Others
TITLE The Effect of Review on Student Ability and Test Efficiency for Computer Adaptive Tests.
PUB DATE Apr 91
NOTE 19p.; Paper presented at the Annual Meeting of the National Council of Measurement in Education (Chicago, IL, April 3-6, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Ability; *Adaptive Testing; Certification; Comparative Testing; *Computer Assisted Testing; *Item Response Theory; Mastery Tests; *Medical Students; Pass Fail Grading; *Review (Reexamination); Test Validity
IDENTIFIERS Rasch Model

ABSTRACT

This paper explores the effect of reviewing items and altering responses on the efficiency of computer adaptive tests (CATs) and the resultant ability measures of examinees. Subjects included 712 medical students: 220 subjects were randomly assigned to the review condition; 492 were randomly assigned to a review control condition (the usual CAT administration method). The test, designed to establish student ability above or below a pass/fail point, included items from a pool of 726 Rasch calibrated items. Subjects in the review condition were instructed to answer each item when presented, but were allowed to review and alter responses at the end of the test. Subjects in the control condition were not permitted to review and alter responses. Student ability measures before and after review were correlated at 0.98. The average efficiency of the test was decreased by 1% after review. About 32% of the students improved slightly after review, but did not change their pass/fail status. These data do not support prohibition of review on CATs administered under these rules. Results indicate that the students allowed to review performed significantly, if slightly, better than did control students. Three tables are included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED330722

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Mary Lunz

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Effect of Review on Student Ability and Test Efficiency for Computer Adaptive Tests

Mary E. Lunz

Betty A. Bergstrom

American Society of Clinical Pathologists

Benjamin D. Wright

University of Chicago

BEST COPY AVAILABLE

Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL, April, 1991.

TM016318

The Effect of Review on Student Ability and Test Efficiency for Computer Adaptive Tests

Abstract

This study explores the effect of reviewing items and altering responses on the efficiency of computer adaptive tests and the resultant ability measures of examinees. Two-hundred-twenty of the 712 students were randomly assigned to a review condition. Their instructions indicated that they must answer each item when presented but that they could review and alter responses at the end of the test. The other 492 students did not have the opportunity to review and alter responses. Student ability measures before and after review were correlated at .98. The average efficiency of the test was decreased by one percent after review. Approximately 32% of the students improved slightly after review, but did not change their pass/fail status. Disallowing review on adaptive tests administered under these rules is not supported by these data.

The Effect of Review on Student Ability and Test Efficiency for Computer Adaptive Tests

The purpose of educational measurement is to inform educational decision making by providing estimates of an individual's knowledge and skill. For certification and licensure, this means making minimum competency pass/fail decisions. In recent years, computers have become more versatile and more accepted for the development and delivery of examinations. One of the most interesting and potentially advantageous methods for test developers and examinees is computer adaptive testing (CAT). The algorithms for item selection usually depend on item response theory (IRT) (Rasch 1960/1980, Lord and Novick, 1968; Wright and Stone, 1979). The items in the bank are calibrated to a common scale on which the pass/fail point has been established. The adaptive algorithm selects items that provide the most information about examinee ability given the current ability estimate from responses to the previous items. Because the items administered are tailored to the performance of the examinee, fewer items are needed to reach a decision with the specified level of confidence. Computer adaptive tests frequently vary in length because the stopping rules require a specified level of precision (measurement error) or level of confidence (distance from the pass point) in the accuracy of the decision rather than a set number of items as in paper and pencil tests.

The usual computer adaptive testing administration procedure is to select a starting point, at the middle of the scale or at the pass point and

to present an item of that difficulty. If the examinee answers the item correctly, a more difficult item is presented. If the examinee answers that item incorrectly, an easier item is presented. Items with difficulties near the current ability estimate of the examinee continue to be presented, so that maximum information is gained from each item (Weiss and Kingsbury, 1984) until the stopping rule is met.

When the difficulty of an item matches the ability of an examinee, he has a 50% probability of answering correctly. When the item is targeted slightly below the ability of the examinee, he has a greater than 50% probability of answering correctly. A stopping rule based on confidence in the accuracy of the pass/fail decision requires that an examinee's ability measure be a specified interval above or below the pass/fail point, usually 1.3 to 2 standard errors (90% to 95% confidence) A confidence level stopping rule produces more precise ability estimates for examinees with estimated abilities near the pass/fail point because they take longer tests. Examinees with abilities distant from the pass/fail point take shorter tests and consequently have larger errors even though their pass/fail decisions are made with the same level of confidence (e.g. 90% or 95%). When a confidence level stopping rule is used, a variable length test with an established minimum number of items is most effective. Each examinee takes a test that measures their ability to the level of precision needed to decide whether they have passed or failed.

The opportunity to review items and alter responses has not been allowed when tests are administered adaptively. One concern has been that reviewing and altering item responses will change the estimate of examinee ability so much that the sequence of items will become poorly targeted and a substantial

amount of precision lost. This concern assumes that the response alterations made by the examinee will substantially increase the distance between the current examinee ability estimate and the previously selected item difficulty. This assumption has not been verified with real data.

Some researchers argue that the precision of the measure must be reduced by response alterations because any change in responses will alter the information value for the next item, and hence for all subsequent items. For example, if an examinee changes two sequential incorrect responses to correct responses, the estimated current ability measure would increase and all subsequent items could be targeted on the "too easy" side, thus reducing the information about the examinee gained from the subsequent items. However, systematic increase or decrease in ability as a result of item response review has not, in fact, actually been observed.

The opportunity to review items and alter responses is usually important to the examinees. Most examinees feel at a disadvantage when they cannot review and alter their responses. In addition, there is the fear of entering an incorrect answer accidentally and not being able to change it. The security of a final review can provide comfort and reassurance to the examinee who is taking a high stakes test.

Allowing examinees to review a computer adaptive test also has implications for the perceived fairness of the examination process. An examination appears more fair to the examinee when review and alteration of responses is allowed. Examinees have one final opportunity to perfect the demonstration of their knowledge by checking for entry errors and "double checking" uncertain responses.

But the effect of reviewing and altering responses on the results of

computer adaptive tests has not, as yet, been studied in any detail. Wise, Barnes, Harvey and Plake (1989) found no significant differences in the mean measures of groups of examinees in review and non-review conditions. But the number of responses altered and the direction of the change (correct to incorrect; incorrect to correct or incorrect to incorrect) has not been studied. If patterns of response during review improve or reduce ability measures systematically, the information value of subsequent items could be altered and the pass/fail decision could become less precise.

This paper explores the effect of reviewing items and altering responses on the ability measure of examinees and the efficiency of computer adaptive testing. Efficiency refers to the amount of information gained from each item administered. The hypothesis is that the opportunity to review items and alter responses will not alter significantly either the efficiency of the test or the examinee measures and resulting pass/fail decisions. The information value of the items presented by the computer adaptive algorithm will remain substantially unaffected by response alterations.

Methods

Test Description

Each examinee took an individualized computer adaptive test selected from a pool of 726 Rasch calibrated items. The test was designed as an adaptive mastery test (Weiss and Kingsbury, 1984) to determine whether an examinee's estimated ability is above or below a pre-established pass/fail point.

The test was administered by the CAT Administrator (Gershon, 1989) which implemented the PROX version of the maximum likelihood method of item selection (Wright, 1977; Wright and Stone, 1979) in the adaptive algorithm.

The calibrated item pool contained items from six subtests. The distribution of item difficulties within each subtest were comparable and subtests were not significantly different in average difficulty. Content coverage followed the test plan for the paper and pencil examinations. Subtest order was selected randomly. Maurelli and Weiss (1983) found subtest order to have no effect on the psychometric properties of an existent achievement test battery.

The minimum number of items was set at 50 and covered all six content areas according to the established test plan. The number of items administered varied, but the maximum was 240. The stopping rule required the measured ability of examinees to be 1.3 times the standard error of measure (Wright and Masters, 1982) above or below the pass point before testing stopped. This rule provides 90 percent (one-tailed test) confidence in the pass/fail decision.

Test Conditions

Two test conditions were used, a "review" condition and a "non-review" condition. Examinees in both conditions took a computer adaptive test which required them to answer each item before another item was presented. The review condition, however, allowed students to review questions and alter responses at the end of the test. Items were presented until the test specifications were completed and a decision about the examinee could be made or 240 items had been administered. Instructions for the review procedure were then presented. During review, items were presented in the same order as they appeared in the test and the student's initial response was highlighted. Students could devote as much time to reviewing and changing answers as they wished. The directions for the non-review condition specified that students must answer each item when presented. Students were cautioned to consider

their answers carefully since they would have only one opportunity to answer the question.

Seven-hundred-twelve (712) students from across the country participated. A random sample of 220 of these students were placed in the review condition. The remaining 492 students were in the non-review condition (usual CAT administration method). Students were serious about this testing because it was part of their review for their certification examination.

Two test records were maintained for each student in the review condition, one before and one after review. The first record contained the item sequence numbers, item difficulties, examinee responses, current ability measures and errors for the computer adaptive test before review. The second record contained the same information after review. Comparison of these records shows the number and accuracy of response changes made during review. From these records, item difficulties and current ability measures can be compared to ascertain the loss of information and hence loss of test precision caused by response alteration.

Test Efficiency

Test efficiency is determined by the amount of information gained from each item administered. Maximally efficient tests produce the most information about the ability of an examinee with the smallest number of items. The efficiency of computer adaptive tests depends upon targeting items to the examinee's ability so that a pass/fail decision at the specified level of confidence is reached with fewer items. Off-target items provide less information and therefore make longer tests necessary.

Calculation of relative test efficiency depends upon the following.

b = examinee ability measure

d = item difficulty

p = (exp b-d)/[1+exp(b-d)] = probability of correctly responding to an item

q = p(1-p) information value of one item

SEM = $(1/\Sigma q)^{\frac{1}{2}}$ standard error of measure

I = Σq = test information value

To analyze the impact of review, the components of test efficiency are applied as follows.

b_1 = ability measure before review

b_2 = ability measure after review

b_2-b_1 = amount of change in measured ability due to review

SEM_1 = standard error measure before review

SEM_2 = standard error of measure after review

SEM^2_2/SEM^2_1 = relative efficiency of the test after review

I_1 = test information before review

I_2 = test information after review

I_2-I_1 = loss of information due to review

$N_{12} = 4*(I_2-I_1)$ = number of perfectly targeted items required to recover information loss

The information gained from an item depends upon the probability of a correct response. The test information (I) is the sum of the information contributed by the items ($I=\Sigma q$). As each item is administered more information is acquired about the student. If a substantial number of items

become "off target" as a result of changes in the examinee's current measured ability due to altering responses during review, the information value may change ($I_2 - I_1$). When an item is perfectly targeted, an examinee has a .5 probability of a correct response. Perfectly targeted items have an information value of .25 logits [(q=p(1-p)]. Thus the number of additional items which would have to be presented to an examinee to compensate for information loss can be calculated as $4*(I_2-I_1)$.

The ratio of the squared measurement errors before and after review indicates the relative test efficiency after review. Efficiency depends on the number of items near enough to the person's ability to make a tangible contribution to the estimated measure.

Results

Comparison Between Groups

The mean ability measures and standard deviations for the review and non-review groups were .24 (.48) and .16 (.50) logits respectively. These means are significantly different ($t=-2.08$, df. 710, $p<.04$). The examinees who were allowed to review performed significantly, but only slightly (0.08 logits) better on average, than the examinees who were not allowed to review.

Review Group Results

Table 1 summarizes the response alterations during review. The minimum number of items altered was 0, the maximum was 16, the average was 2. 85 examinees did not alter responses even though the opportunity to do so was available. 135 examinees altered some responses during review. Of these examinees, 30 earned lower measures after review, 71 earned higher measures after review, and 34 effected no change in their measures as a result of altering responses. The examinee who altered 16 responses changed 9 items

from incorrect to correct, 6 from correct to incorrect and 1 from incorrect to incorrect. His test contained 95 items, so 17% of the items were altered. Overall, the percentage of item responses altered during review was very small (2%). The patterns of alteration did not usually involve changes to a series of sequential items. Rather one or two item responses were altered at diverse points in the test (e.g. items 3, 15, 27).

INSERT TABLE 1 ABOUT HERE

Table 2 presents the summary of ability measures and test efficiency for the review group. The mean measure and SD before review (b_1) was .22 (.47) and after review (b_2) was .24 (.48). Mean change was .03 logits. Examinee measures before and after review correlated at .98.

INSERT TABLE 2 ABOUT HERE

The relative test efficiency ratio was 99% with a standard deviation of 3%. Thus the average efficiency of the test decreased by one percent after review. Information loss (I_2-I_1) due to review was .11 logits. On average, the information loss can be recovered by the addition of one-half item (N_{12}). Of the 135 students who altered responses, 108 would require no additional items, 23 would require 2 to 5 additional items, and 4 examinees would require 6 to 14 additional items to recover the information loss.

The overall effect of altering responses on the pass/fail decisions was also minimal. Table 3 shows the tabulation for the review group. Three examinees altered their status from fail to pass after review. These examinee measures were within an error of measure of the pass/fail point both before and after review, indicating less than 70% confidence in the accuracy of either decision. The average improvement for these 3 examinees was .05 logits, the result of changing one response from incorrect to correct.

INSERT TABLE 3 ABOUT HERE

Discussion

There are two issues involved in allowing examinees taking a computer adaptive test to review items and alter responses. The first issue concerns the psychometric implications of including items that provide less than maximum information about the examinee. When the response to an item previously answered correctly or incorrectly is altered, the examinee measure and the standard error are altered, yet the difficulty of the next item administered remains the same. The question is how close to the revised ability estimate must the item difficulty be to maintain the efficiency of the item selection algorithm. In these data, the difficulties of subsequent items were within the related error of measure of the revised ability estimate, even though some items were altered from incorrect to correct or vice-versa. The mean loss of information due to these changes was less than the amount of information provided by one additional on-target item. The one exception is the single examinee who altered 14 items, 10 from incorrect to correct, thus significantly improving his ability measure. His status, however did not

change, it was pass to pass.

The second issue is the face validity of computer adaptive testing. Educators have trained students from elementary school to medical school to review responses after completing a test to catch careless errors and thus render a more accurate demonstration of competence. When adults, who grew up in this system, face a certification examination, they usually feel it is their "right" to review their work for careless errors. Indeed, the testing goal is not to penalize an examinee for clerical errors in the name of psychometric integrity, but rather to measure their ability as accurately as possible. These data show that decision accuracy and confidence are comparable before and after review. The opportunity to review and alter responses, did not significantly change these examinee ability measures. Review can, however, give examinees confidence in the system and a sense of having made every effort to demonstrate accurately their knowledge and skill.

The limitation of this study is that the examinees knew that this was not an actual certification examination. They were encouraged to perform their best, but there was no immediate reward attached to acceptable performance. This might not, therefore, be typical examinee review behavior on the certification examination. Examinees may alter more or perhaps even fewer responses. The next phase of this study involves a direct reward for acceptable performance (e.g. certification). Results from the next phase will provide more insight into the way examinees use the opportunity to review.

The fact that the group allowed to review performed significantly, if ever so slightly (0.08 logits), better than the randomly equivalent group that was not allowed to review encourages the "greater confidence" theory. It is easy to suppose that the reason for the difference relates to the familiarity

and comfort associated with the knowledge that review and correction of careless errors was allowed. The response alterations result in measure changes well within the error of measurement. The personal control associated with the opportunity to review may inspire more careful consideration of each item leading to overall improved scores.

References

Gershon, R. (1989). CAT Administrator. Chicago: Micro Connections.

Koch, W.R. and Dodd, B.G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. Applied Measurement in Education, 2, 335-357.

Lord, F.M. (1977). A broad range tailored test of verbal ability. Applied Psychological Measurement, 1, 95-100.

Lord, F.M. and Novick, M.R. (1968). Statistical Theories or Mental Test Score. Reading, MA: Addison-Wesley.

Lunz, M.E., Bergstrom, B.A., Gershon, R. and Wright, B.D. (1989). Test/Retest consistency of computer adaptive test. Paper presented in the annual meeting of the NCME, Boston, MA.

Maurelli, V.A. and Weiss, D.J. (1983). Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries (Research Report 81-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Olsen, J.B., Maynes, D.D., Slawson, D. and Ho, K. (1986). Comparison and equating of paper-administered, computer-administered and computerized adaptive tests of achievement. Paper presented at the American Educational Research Association Meeting, San Francisco, CA.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Weiss, D.L. and Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21, 4, 361-375.

Wise, S.L., Barnes, L.B., Harvey, A.L. & Plake, B.S. (1989). Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. Applied Measurement in Education, 2, 235-241.

Wright, B.D. and Stone, M. (1979). Best Test Design. Chicago, IL: MESA.

Wright, B.D. and Masters, G. (1982). Rating Scale Analysis. Chicago, IL: MESA

Wright, B.D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

TABLE 1

Pattern of Response Alterations
During Review
N=220 Examinees

Descriptive Statistics		
Total N of responses =	21,082	
N of responses altered =	456	(2%)
Mean N of items taken =	96	
Mean N of item altered =	2	

Response Alteration	N of Responses Altered	Percent
Incorrect to correct	214	47%
Correct to incorrect	123	23%
Incorrect to incorrect	119	26%
Total	456	100%

TABLE 2

Summary of Ability Measures and Relative Test Efficiency Before and After Review

N=220 Examinees

	Mean	SD
Ability Measurement*		
B_1	.22	.47
B_2	.24	.48
$B_2 - B_1$	.03	.09
Relative Test EFficiency		
SEM_1	.23	.05
SEM_2	.23	.05
SEM^2_2/SEM^1_2	.99	.03
Information Value		
I_1	22.58	12.48
I_2	22.46	12.55
$I_2 - I_1$	-0.11	.51
Number of Items to Recover Information		
N_{12}	-.45	2.03

*Reported in logits

TABLE 3

Pass/Fail Decisions
Before and After Review

		BEFORE: Pass	BEFORE: Fail
AFTER	Pass	120	3
	Fail	0	97

SAME DECISION FOR 217 (99%)

Mean improvement for Pass/Pass = .04 logits
Mean improvement for Fail/Fail = .01 logits
Mean improvement for Fail/Pass = .05 logits

Mean improvement overall = .03 logits